MOTION INTEGRATION MODULATED BY FORM INFORMATION

Émilien Tlapale
Équipe Projet Odyssée
INRIA
Emilien.Tlapale@inria.fr

Guillaume S. Masson
Équipe DyVA
INCM - CNRS
guillaume.masson@incm.cnrs-mrs.fr

Pierre Kornprobst
Équipe Projet Odyssée
INRIA
Pierre.Kornprobst@inria.fr

ABSTRACT
We propose a model of motion integration modulated by form information, inspired by neurobiological data. Our dynamical system models several key features of the motion processing stream in primate visual cortex. Thanks to a multi-layer architecture incorporating both feedforward-feedback and inhibitive lateral connections, our model is able to solve local motion ambiguities. The main feature of our model is to propose an anisotropic integration of motion based on the form modulation. The proposed mechanism is not only simple but is also not limited to a fixed number of depth/scale layers [1] and does not blindly detect and ignore all junctions [2, 3]. Our model can be implemented efficiently on GPU and we show its properties on classical psychophysical examples. First, a simple read-out allows us to reproduce the dynamics of eye movements for a moving bar stimulus. Second, we show how our model is able to discriminate between extrinsic and intrinsic junctions present in the chopstick and Lorenceau-Alais [4] illusions. We also show how our form modulation induces a notion of objects explaining recent experiments [5]. Finally, we show some promising results on complex and real videos.

KEY WORDS
motion integration, feedbacks, GPU, motion perception, extrinsic junctions

1 Introduction

Many bio-inspired models of motion processing by the visual cortex of primates have been proposed to explain the complex motion integration mechanisms and the psychophysical experiments. We can distinguish two big classes of approaches. The first class of approaches is primarily high level and their main goal is not to focus on the precise anatomical or functional properties the visual system but to show of some fundamental principles that permit to reproduce some psychophysical results (see, e.g., the Bayesian models [2, 6]). The second class of approaches aims at modeling some of the key features of the visual system in term of structure and connectivity, and to show how this so-called bio-inspiration allows psychophysical effects to be reproduced. The model we propose belongs to that later class and it is inspired by some recent contribution [7, 8, 1].

In order to compute the global motion of a scene, motion processing systems, and similarly the visual cortex, take local motion estimates as input. The problem is that this local motion information, which is estimated over a limited spatial neighborhood, is in general noisy and ambiguous, leading to the well known *aperture problem* [9]. For example, along contours only motion perpendicular to the contour can be perceived, a problem illustrated in Figure 1. To solve the aperture problem, we need to integrate motion information from other non ambiguous areas like corners.

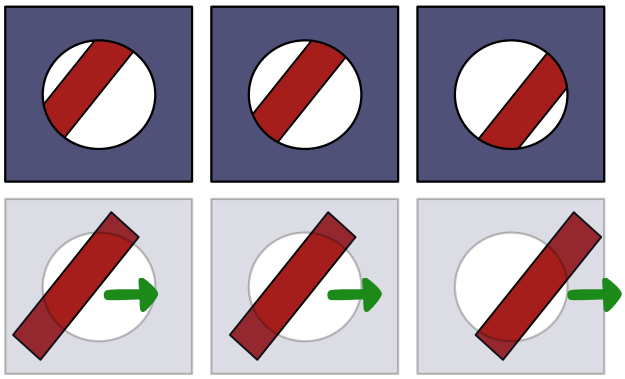


Figure 1. Seeing a translating bar through a circular aperture masking its ends does not give enough information to know its real motion, we only perceive a motion orthogonal to the bar direction. Displaying the bar ends we are able to spatially integrate information and perceive the correct motion.

Diffusion of local features, or equivalently regularity constraints, have been proposed to solve the aperture problem in machine vision. Interestingly, the visual system also performs such a diffusion by an integration of the local motion features across different layers: motion signals from unambiguous regions such as line endings are propagated inside the moving objects. It is this motion diffusion process based on integration rules which solve the aperture problem. For instance [1] use multiple rules based on junction detectors, depth/scale layers, and motion direction.

A *winner-take-all* mechanism can be used in order to select and enhance the motion signals to be diffused. Indeed the raw signal from motion detectors is often noisy and ambiguous. Various bio-inspired approaches can be

found in the literature for signal amplification, for instance in feedforward systems [7] and using divisive [10] or subtractive [11] inhibition. After local amplification, unambiguous motion signals may be diffused spatially to help disambiguation of areas affected by the aperture problem. This diffusion may either be done thanks to local intracortical connections or via feedbacks from other cortical areas [8].

Incorporating more features, in particular form information, help to obtain more accurate results. Recent models employ this strategy to get a better output in term of velocity estimation and of psychophysical results. Indeed, motion models compute motion perception from locally unambiguous motions not affected by the aperture problem. However psychophysical studies [12] shows that some of the unambiguous motion signals are ignored in global motion perception. Those studies thus classify locally unambiguous signals as either *intrinsic junctions* if they are globally integrated, or *extrinsic junctions* if they are ignored in the global percept. We later describe the chopstick illusions, a classical example of the importance of extrinsic and intrinsic junctions. In order to discriminate between extrinsic and intrinsic motion signals, some models discard certain kind of junctions [13, 3] assuming their extrinsic probability based on their geometrical characteristics. Other models segregate motion information into two layers according to form features [1] thus allowing a limited transparent motion process.

In this paper we propose a new motion integration mechanism based on a directional motion diffusion which is form modulated. Section 2 describes the model and its biological interpretation. Section 3 presents some of the results obtained for both synthetic and real sequences. We conclude in Section 4.

2 Towards a model of motion integration

2.1 Coupled dynamical systems

Our model describes the activity and the interactions between different layers as a coupled dynamical system. This widely adopted formalism [8, 1] reflects the brain division in *cortical areas*. The state of a layer i is defined in our model by the function

$$p_i : (t, x, v) \in \mathbb{R}^+ \times \Omega \times \mathcal{V} \rightarrow p_i(t, x, v) \in [0, 1]. \quad (1)$$

where t is the time, $x = (x_1, x_2)$ denotes the spatial position belonging to the 2D-spatial domain $\Omega \subset \mathbb{R}^2$, and $\mathcal{V}$ represents the space of possible velocities. This function p_i can be interpreted as the state of a cortical area retinotopically organized which describes at each position the instantaneous activity of a velocity tuned neuron.

Since we implement a multi-layer system, we will describe the evolution of the activity in each layer in respect with the activity of the whole system. The dynamical description leads to a differential equation in time which combine a local decay and various interactions:

$$\dot{p}_i(t, x, v) = -\lambda_i p_i(t, x, v) + f_i(p_0(t, x, v), p_1(t, x, v), \ldots). \quad (2)$$

a description also employed by [1] and [8], even if the later do not use a dynamical system in the implementation. Following equations do not explicit the (t, x, v) parameters of the layers.

2.2 Proposed model

The initial stage of every motion processing model is to compute local motions cues, denoted by $p_0(t, x, v)$, from an input video sequence $I : (t, x) \mapsto I(t, x) \in \mathbb{R}$. In the visual system, this local processing is done at different levels and time, from the retina to the visual cortex. Some existing approaches, such as [7], model this process with banks of spatio-temporal filters. In this paper, we will use the implementation of the Reichardt detectors of [8], which estimates some correlations between delayed filters at neighboring areas.

Our model is defined by the interaction of two coupled cortical layers, p_1 and p_2, depicted in Figure 2 and defined by:

$$\dot{p}_1 = -\lambda_1 p_1 + \quad (3)$$
$$(1 - p_1)\big[\lambda_a p_0 + \lambda_b p_0 p_2 - \lambda_c G_{\sigma_1} \overset{x}{*} \int_V p_1(t, x, w) \mathrm{d}w\big]_+$$
$$\dot{p}_2 = -\lambda_2 p_2 + \quad (4)$$
$$(1 - p_2)\big[\lambda_m G_{\sigma_2} \overset{x}{*} \int_\Omega G_{\sigma_x}(x - y)\phi(t, y, \widehat{yx}) p_1(t, y, v) \mathrm{d}y$$
$$- \lambda_n G_{\sigma_2} \overset{x}{*} \int_V p_2(t, x, w) \mathrm{d}w\big]_+.$$

where $\dot{p}_i = \frac{\partial p_i}{\partial t}$ is the partial derivative in time of p_i, $\widehat{yx}$ denotes the angle of the vector yx in retinotopic coordinates, $[\cdot]_+$ is the rectification operator defined by $[s]_+ = \max(0, s)$, the $\lambda_\bullet$, $\sigma_\bullet$ are constants.

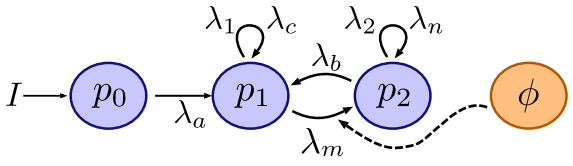


Figure 2. Schematic view of the proposed model showing the interactions of the different cortical layers. The motion integration (p_0, p_1 and p_2) system is modulated (*dashed arrow*) by a form information (ϕ).

The evolution of the two main layers, p_1 and p_2 is defined by differential equations, characterizing their behavior across time. This model is inspired by [8], where the

authors make the correspondence between these layers and layers V1 and MT. Both layers contain a leak ($-p_i$) which stabilizes the system by attracting the state to zero. Then, the factor $(1-p_i)$ has been chosen in order to constrain the activation rate to be in the interval $[0, 1]$.

2.3 Model features

Feedback integration Our first layer,p_1, combines feedforward input from p_0 and feedback from the second layer, i.e., p_2. This structure is inspired by [8]. To allow motion diffusion and integration p_2 neurons have access to multiple p_1 neurons in an anisotropic neighborhood. This follows observations concerning the increase of receptive field sizes from V1 to MT. Note that the feedback from p_2 are combined in a multiplicative way in p_1 as in [8] supporting the *no strong loop hypothesis*: feedback alone cannot evoke a response in our system.

Form-modulated diffusion In layer p_2, we integrate motion information from p_1 in a spatial neighborhood. This spatial neighborhood is not defined by a simple isotropic and invariant Gaussian smoothing, it also depends on the input stimulus through form information which is processed in area V2. V2 neurons can extract edges/shape information from different cues (i.e. luminance, relative motion, disparity, ...). The role of shape in general has been demonstrated in several psychophysical experiments [12, 5].

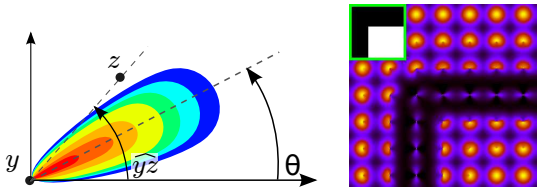
Here we propose to use shape descriptors *positively*, i.e., to control a diffusion instead of suppressing it in the presence of complex structures. To do this, we define a shape function, ϕ, which can be related to V2 cells [14] and defined as:

$$\begin{aligned} &\phi : \mathbb{R}^+ \times \Omega \times [0, 2\pi[\rightarrow \mathbb{R}^+ \\ &\phi(t, y, \theta) = \int_\Omega w(y, z, \theta) G_{\sigma_s}(I(y) - I(z)) \mathrm{d}z \qquad (5) \\ &\text{where } w(y, z, \theta) = G_{\sigma_x}(y - z)\, G_{\sigma_\theta}(\theta - \widehat{yz}) \end{aligned}$$

Equation (5) describes the power of diffusion at a given point y and in a given direction θ. In this article we only consider luminosity information $I(t, y)$ as form information. Thus we diffusion information from a point y in a direction θ if the luminosity in this direction, i.e. in the neighborhood w, is similar to the one at y. In Figure 3 we display the directional neighborhood w and a sampled representation of ϕ for an orthogonal edge frame.

Thanks to the ϕ function, the stimulus dependent integration process has two main properties that have been observed in cell recordings: integration is facilitated inside similar structures (see, e.g., [5]) and the extension of the integration also depends on the local contrasts (see [15]).

Lateral inhibition The last part of the equations defining our layers is the lateral inhibition. All neurons at a given local neighborhood for all possible velocities inhibits one the other. This lateral inhibition, sometimes called recurrent inhibition, leads to a winner-take-all mechanism [11]. Instead of this kind of subtractive inhibition, a divisive inhibition has also been successfully used [10, 8].



(a) (b)

Figure 3. (a) To compute $\phi(t, y, \theta)$ we integrate the differences of luminosity between $I(t, y)$ and the points in a spatial weighted neighborhood, $w(y, z, \theta)$. (b) Spatially sampled representation of ϕ for a synthetic corner displayed in the green square. At each point of the sampling the weight of diffusion along each direction is displayed.

3 Results

Material and methods A discretization procedure has to be applied since we work on dynamical equation: we choose the Runge-Kutta algorithm. Moreover since the input is not continuous but is made of successive frames, and because we want more precision than the coarse input, we need intermediary frames. For simplicity, we did not interpolate but choose input similar to the previous frame for all intermediate frames before the next one. We discretized the system with ten intermediary time steps between two input frames, not including the intermediary frames of Runge-Kutta.

On the following results we choose our discretized velocity space to be all the integer pairs $v = (v_x, v_y)$ in a 7×7 regularly spaced grid. As a read-out, we can extract one velocity field $\mathrm{v}_i(t, x)$, i.e. a single motion at each spatial position, for the layer p_i by:

$$\mathrm{v}_i(t, x) = \big(\sum_v p_i(t, x, v)\, v\big) / \big(\sum_v p_i(t, x, v)\big). \qquad (6)$$

For the calculation of ϕ we used the following parameters: $\sigma_x = 12$, $\sigma_\theta = \pi/8$, $\sigma_s = 0.4$. We fixed motion integration parameters to $\lambda_1 = \lambda_2 = 4$, $\lambda_a = 1$, $\lambda_b = \lambda_m = 16$, $\lambda_c = \lambda_n = 4$ and Gaussian radius to $\sigma_x = 10$, $\sigma_1 = 4$, $\sigma_2 = 8$.

Because of the anisotropic diffusion depending on input stimulus, our model takes a considerable computational effort. Conventional CPU implementation is far from being fast, so we implemented our model on GPU to take advantage of its parallel nature using NVIDIA's CUDA technology. Except for GPU kernels, all the code is written in Python using the SciPy library.

Motion integration The motion integration and disambiguation mechanisms can be illustrated with a translating bar (see Figure 4 (a)). Figures 4 (b)-(d) display the velocity field computed at the first iteration and after some iterations of our model according to Equation (6). The colormap of Figure 4 (e) is used to associate a color to each velocity. Note how the end of line information is propagated towards the center of the bar.

In Figure 4 (f) we display our *read-out* computed by averaging the velocity field over the whole stimulus for each frame. It can be associated to the eye movements and indeed show a shape similar to what can be found in psychophysical literature for the same stimulus [16].

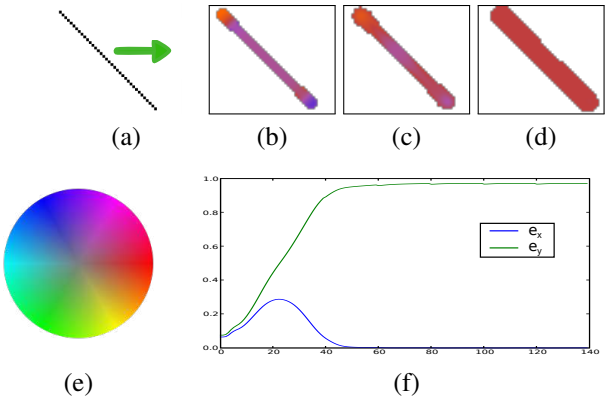


Figure 4. Response of the model on a horizontally translating bar presented in (a). (c)-(d) Evolution of the velocity field $\mathrm{v}_1(t, x)$. (e) Color code used for the velocity fields. (f) Temporal read-out providing a global motion similar to the one get in eye movements computed by averaging the velocity field. Green and blue correspond respectively to the v_x and v_y components.

Extrinsic/intrinsic junctions We use the chopstick illusions to illustrate the influence of form information in motion processing. The first stimulus is made of two horizontally translating bars (see first line of Figure 5). We thus have unambiguous motion information from the end of lines, the horizontal motion, and from the bars intersection, the vertical motion. We display the velocity field v_1 and show that our results are coherent to psychophysical experiments where two horizontal bars are perceived.

In the second line of Figure 5 we use the same stimulus with two rectangular occluders at the the end of lines level. Again our results are coherent with psychophysical experiments where one vertical motion is perceived. In both experiments we use the same stimulus characteristics.

The same simple form modulated motion integration model has also been applied to Lorenceau-Alais illusions [4], see Figure 6. We obtain results similar to [1]: motion information compatible with the correct rotation motion in the diamond case but two translational motions in the arrow case; similar to psychophysical results. Again, without relying on depth/scale layers, neither on junction detectors, nor on complex rules, as in [1], but only on our directional form information we are able to reproduce human motion percept.

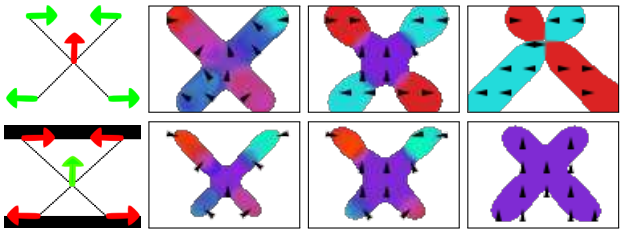


Figure 5. The *first line* show the non-occluded chopstick illusion made of two horizontally translating bars. For perception, the end of line 2D information is propagated (*green*), the intersection 2D information is inhibited (*red*). We display the velocity field v_2 obtained from our model which exhibit the same behavior, i.e. having two horizontal motions. On the *second line* we use the occluded chopstick illusion which adds two rectangular occluders at the end of lines level. This change the perception to a single vertical motion. A result which is also reproduced by our model.

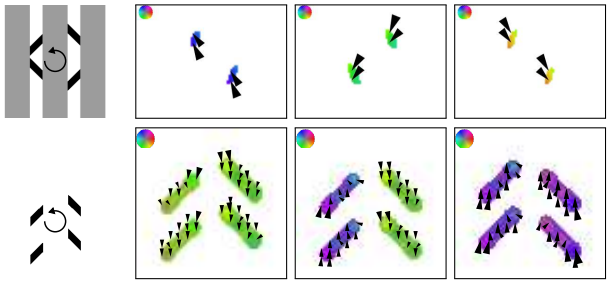


Figure 6. The *two first rows* display the output of our model (v_1) applied to Lorenceau-Alais diamond illusion [4] for three different frames. We observe a percept coherent with rotation with results similar to the model in [1]. The *two last rows* display the output or our model (v_1) for the arrow Lorenceau-Alais illusion [4]. We observe two translational motions instead of the rotation like in [1].

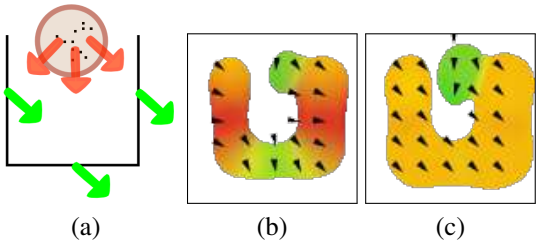
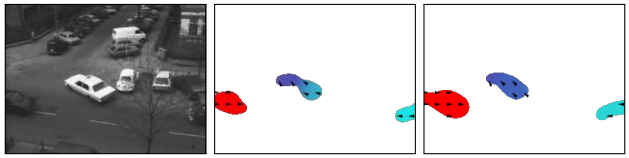


Figure 7. (a) Square moving left and down with points moving randomly down at top removed edge (see [5]). (b) Velocity field in p_2 at the beginning with the aperture problem. (c) Iterating does not solve the aperture problem on top.

Diffusion on objects

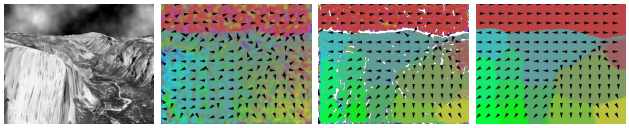
In Figure 7 we use the stimulus used in [5]: a square moving in the lower right direction with its top edge removed and replaced by a set of points moving randomly downward. The points reproduce the velocity distribution in the aperture problem at the center of an edge. Our model gives results similar to the cells recording: the ambiguity is not solved in the replaced edge and the velocity field is thus averaged as a downward motion.

Complex and real sequences We also applied our motion processing model on real sequences, such as the Taxi sequence. Results for this sequence are shown in Figure 8 displaying the segmentation of moving objects in homogeneous regions by our method.



(a) (b) (c)

Figure 8. Motion field obtained by processing the Taxi video sequence in our model. (a) A frame of the used video sequence. (b) Initial velocity in p_2. (c) Velocity field after a few frames.



(a) (b) (c) (d)

Figure 9. (a) A frame of the Yosemite sequence. (b)-(d) Velocity field computed respectively from p_0, p_1 and p_2

In Figure 9 we display the velocity field computed from our model on the Yosemite video sequence. All motion processing layers are shown: p_0, p_1 and p_2 for a given time. Note the *patch effect* due to the limited range of velocities and the winner-take-all nature of our system.

4 Conclusion

We described a motion integration mechanism which disambiguate local motion signal by incorporating a form information into the pure motion system. By doing so we are able to discriminate between intrinsic and extrinsic junctions leading to motion percept coherent with the psychophysic, as illustrated with the chopstick illusion, or the diamond illusions [4]. Moreover our model is able to discriminate between moving objects and segment them segment moving object in real video sequences, again without the need of explicit junction detectors [13, 3] or motion layers [1].

The dynamical computation of our model enables us to compute a simple read-out representing the velocity of the object of interest. Such output can be compared to the temporal dynamics of smooth pursuit eye movements in humans [16] where the tracking direction errors closely match the estimated 2D velocity. Such comparison would be more difficult using a coarser frame-by-frame algorithm [3].

Our model however is not yet able to perceive transparent motion directly despite the use of a distributed motion representation similar to the one found in the visual cortex. Additionally certain stimuli may need a more sophisticated form modulation even if the one presented is able to discriminate between extrinsic and intrinsic junctions in the chopstick illusion where a 3D model has been previously suggested as a possible explanation. Yet to be investigated is also the influence of a large scale feedback, as reported by the biological literature, instead of the simpler local one used in the model. Future work may also use weighted lateral inhibition in order to compute more accurate smooth motion fields and remove the patching effect.

Acknowledgments

This work was partially supported by EC ICT – PROJECT No. 215866 - SEARISE and the Région Provence-Alpes-Côte d'Azur.

References

[1] J. Berzhanskaya, S. Grossberg, and E. Mingolla. Laminar cortical dynamics of visual form and motion interactions during coherent object motion perception. *Spatial Vision*, 20(4):337–395, 2007.

[2] Y. Weiss and E. H. Adelson. Slow and smooth: A Bayesian theory for the combination of local motion signals in human vision. *Center for Biological and Computational Learning Paper*, 1998.

[3] P. Bayerl and H. Neumann. Disambiguating visual motion by form-motion interaction—a computational model. *International Journal of Computer Vision*, 72(1):27–45, 2007.

[4] J. Lorenceau and D. Alais. Form constraints in motion binding. *Nature Neuroscience*, 4:745–751, 2001.

[5] X. Huang, T. D. Albright, and G. R. Stoner. Adaptive Surround Modulation in Cortical Area MT. *Neuron*, 53:761–770, March 2007.

[6] A. Montagnini, P. Mamassian, L. Perrinet, E. Castet, and G.S. Masson. Bayesian modeling of dynamic motion integration. *Journal of Physiology-Paris*, 101(1-3):64–77, 2007.

[7] EP Simoncelli and DJ Heeger. Model of Neuronal Responses in Visual Area MT. *Vision Research*, 38(5):743–761, 1998.

[8] P. Bayerl and H. Neumann. Disambiguating Visual Motion Through Contextual Feedback Modulation, 2004.

[9] H. Wallach. Über visuell wahrgenommene bewegungsrichtung. *Psychological Research*, 20(1):325–380, 1935.

[10] S. J. Nowlan and T. J. Sejnowski. Filter selection model for motion segmentation and velocity integration. *J. Opt. Soc. Am. A*, 11(12):3177–3199, 1994.

[11] A.L. Yuille and N.M. Grzywacz. A Winner-Take-All Mechanism Based on Presynaptic Inhibition Feedback. *Neural Computation*, 1(3):334–347, 1989.

[12] S. Shimojo, G. H. Silverman, and K. Nakayama. Occlusion and the solution to the aperture problem for motion. *Vision Res*, 29(5):619–26, 1989.

[13] Y. Weiss. *Bayesian motion estimation and segmentation*. PhD thesis, Massachusetts Institute of Technology, 1998.

[14] J. Hegdé and D.C. Van Essen. Strategies of shape representation in macaque visual area V2. *Visual Neuroscience*, 20(03):313–328, 2003.

[15] C. C. Pack, J. N. Hunter, and R. T. Born. Contrast dependence of suppressive influences in cortical area mt of alert macaque. *J Neurophysiology*, 93(3):1809–1815, Mar 2005.

[16] J.M. Wallace, L.S. Stone, and G.S. Masson. Object Motion Computation for the Initiation of Smooth Pursuit Eye Movements in Humans. *Journal of Neurophysiology*, 93(4):2279–2293, 2005.